International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.10, No. 1 (2017), pp.233-242
http://dx.doi.org/10.14257/ijsip.2017.10.1.23

Multi-Object Tracking Based on HOG Template Matching and Non-maximum Convergence Algorithm

Lin Xiong[1], Xiaofeng ZHANG[2], JianLan Liao[3] and GuoWei Yang[4]

[1,2,3,4]School of Information Engineering, Nanchang Hangkong University, No.696, Fenghe Nan Avenue, Nanchang, Jiangxi Province, 330063,China
optimus1009@163.com

Abstract

Detection method proposed in this paper is based on histogram of oriented gradients features combining with the template matching to improve the accuracy of pedestrian detection.Then we use support vector machine to train them two times in order to improve the accuracy.We can make a number of weak classifiers with different weights to combine a power strong classifier.At each level we use AdaBoost Algorithm to get a strong classifier and connect them together to form a cascade classifier.Its error rate is very low,but training time is longer.In order to reduce computation time,we can use decision tree to improve it. And a Non-maximum algorithm is proposed,This algorithm mapping multi-results into 3D space and discriminating maximum points in the local area and can effectively detect targets,evaluating their tensity and then getting the best one.

Keywords: *HOG,Template Matching,Cascade Classifier,Non-maximum convergence*

1. Introduction

Pedestrian detection is a key problem in machine vision [1-2]. It has important meaning for improving the quality of life in contemporary society, and it is becoming a research hot spot during recent years. For example, intelligent monitoring systems for buildings,vehicle auxiliary systems, motion analysis, advanced man-machine interface application, these areas are popular application areas of pedestrian detection.Since the diversity of environmental detection background, illumination variation, the uncertainty of pedestrian movement and posture make pedestrian detection different from other general target detection.

Currently, the main detection methods of the pedestrian detection is based on histogram of oriented gradients features and based on Haar features [3-4]. Detection method proposed in this paper, is based on histogram of oriented gradients features combining with the template matching to improve the accuracy of pedestrian detection. Because of the uncertainty of pedestra's position, particularly the uncertainty of the four limbs, will greatly influence the detection results, and the pedestrian's head shoulder position has a good invariance regardless of how is the pedestrian's position. So we add the template of human head shoulder to do detection on the head shoulder part of the target. This method can enhance the detection rate.

Support Vector Machine(SVM)[5-6], a simple binary classifier, is widely used. These two elements make the HOG and SVM based human detection algorithm. When there are many regions needed to be computed repeatedly in feature extraction, integration method can be used. It could shorten computing time and circumvent the disadvantage of lacking global information. Meanwhile, cascade performs well in classification. It consists of several levels of SVMs which makes feature match faster but enlarges classifier training time and computation. There are many result fusion methods [7-8]. The algorithm of Non Maximum Suppression [9-10]can fix them by mapping multi-results into 3D space, evaluating their tensity and then getting the best one. The HOG and SVM based human

ISSN: 2005-4254 IJSIP

detection method could detect human of variant poses in complex backgrounds. While the Integrated HOG and Cascade based detection method accelerates feature extraction and classification,while maintaining the accuracy.

2. The Introduction of HOG

2.1. HOG Feature

Initially, HOG algorithm is used to detect pedestrians in static image,the average detection time of this method can not meet the real time, then, as general-graphics processing units (GpGpus) fast parallel computing [11-13], HOG may be used for real-time applications.The following is an HOG algorithm steps.

2.1.1. Gradient Compute: Gradient compute is the first step of HOG algorithm,The Gradient definition is given as:

$$\begin{aligned} G_x(x,y) &= H(x+1,y) - H(x-1,y) \\ G_y(x,y) &= H(x,y+1) - H(x,y-1) \end{aligned} \tag{1}$$

Where H is the input image， G_x is horizontal Gradient and G_y is vertical gradient， In order to compute each pixel gradient,we need to scan each pixel in the image,The Convolution kernel is formulated as:

$$[-1,0,1] \text{ and } [-1,0,1]^T$$

The gradient of point (x,y) is defined as：

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{2}$$

The gradient direction at (x,y) is defined as:

$$\alpha(x,y) = \tan^{-1}\left(\frac{G_x(x,y)}{G_y(x,y)}\right) \tag{3}$$

2.1.2. Unit Histogram: After compute the histogram, we define a fixed size of detection window to scan image, then, the detection window split into some 8*8 pixel group ,these group are known as unit, the shape of a unit can be a rectangle, also can be a circle, and its size is variable. Each pixel's amplitude in each channel defined as:

$$V_k(x,y) = \begin{cases} G(x,y) & \alpha(x,y) \in bin_k \\ 0 & \alpha(x,y) \notin bin_k \end{cases} \qquad 1 \le k \le 9 \tag{4}$$

2.1.3. Descriptor Blocks: In order to input gradient histogram into the classifier,Unit is constructed in the form of a 3x3, it referred to as block, these blocks can help reduce the impact of illumination and contrast. these blocks generate more spatial information in the image, it also improves the overall performance of detection, Figure 1, shows a example,it gives a detection window contains 9 unit blocks.

Figure 1. HOG Detection Window with Cells and Blocks

3. Template Matching

In simple terms, the template matching use predefined shape to make matching with test image to detect targets. Generally, we store the target template first, which can be gray form or profile form, By using detection window to detect image, then we can calculate the distance of image and template, which also can be said similarity, and we consider the minimum distance that is biggest similarity template is the detection.

The process of a template matching, first we transform image to DT(distance transform)image, and then calculate the chamfer distance of a template and DT image, the object of distance transform are generally binary image. First, the distance from every pixel to its nearest contour is calculated. edge pixel gray value is zero, The resulting image is called DT image, This is a distance transformation formula:

$$DT(p) = \min\{d(p,q) \mid p \in I\} \tag{5}$$

Where I is feature point set of the image, p is pixel in image, q is a edge pixel nearest far from p , $d(p,q)$ denotes Euclideam distance of A to B, $DT(p)$ is gray value after transformation.

The next are Chamfer distance, here due to the use of distance transformation to get the Euclideam distance, the image will have a lot of saw-tooth slopes, so we use Chamfer distance, the Chamfer distance is used for measuring the distance between the template and image, according to the thought that the more short of the distance, the more similar among the models,the Chamfer distance modeled by:

$$D(F,DT) = 1/F\{\sum DT(f) \mid f \in F\} \tag{6}$$

Where F is a template , f is the feature point in template, $DF(F)$ is the distance between f and DT image, here we consider the minimum Chamfer distance of the domain is the position that match the template best. But there's a limit condition, there are strict restrictions on the number of feature points, the area of $F \geq$ a threshold to prevent the Chamfer distance is too short.

The head-body ratio for ordinary people usually falls between 6 and 8， but it can reach 9 for some models, so we can not set the limits too low. So we may assume that pedestrian area obtained by HOG features as a $W \times H$ rectangle, and basing on the definition about the components of human body[17],we set the region $0.37H \times 0.75W$.

3.1. The Fusion Detection Result based on HOG Feature and Template Matching

In order to verify the validity of HOG feature in association with a template matching , we use this method proposed in this paper, compared with a pedestrian detection method based on HOG feature and LBP [18] feature on INRIA dataset and the personal dataset , since MIT dataset is relatively simple, the detection effect is very well,

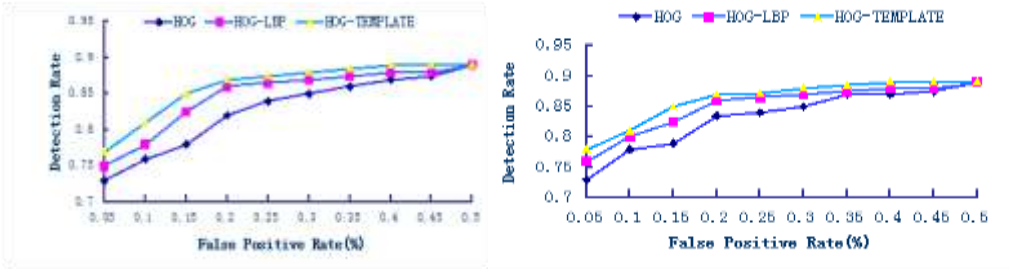
we cannot compare the two methods, so we will use the MIT dataset as the training set,the INRIA dataset and the personal dataset as the test set, Figure(2), Figure(3), and Figure(4), shows some samples.

Figure 2.MIT Dataset

Figure 3. INRIA Dataset

Figure 4. Personal Dataset

Figure (5), gives the ROC curve of three methods on two test sets. Pedestrian detection method based on HOG and LBP feature is better than which based on HOG feature method, while, the proposed method is better than which based on HOG and LBP feature detection method.



a.INRIA Dataset b.Personal Dataset

Figure 5. (a)is the ROC Curve on INRIA Dataset,(b)is the ROC Curve on Personal Dataset

Figure (6), shows the detection result of proposed approach on the test sets, we can see that our algorithm can detect pedestrians at whether single or multi-object scenarios, But also appeared a misjudgment, the reason for this is the complexity influence HOG feature extraction, also because the lack of pixels at head and shoulders, leading edge extraction is not clear.

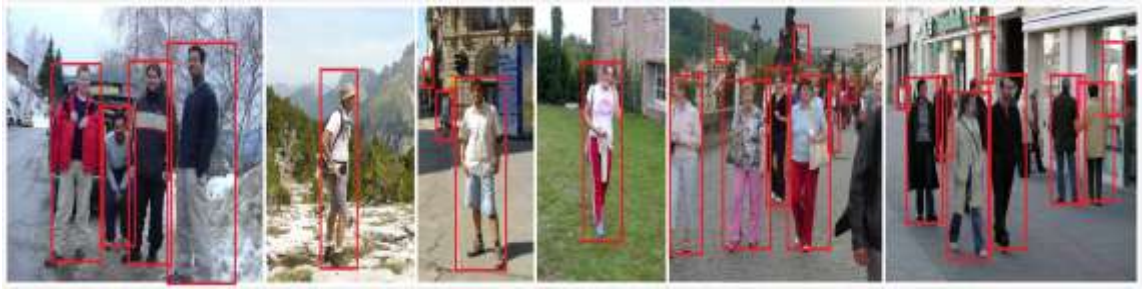


Figure 6. Pedestrian Detection Result

4. Cascade Classifier Training

The Cascade consists of several weak classifiers, Figure (7), each concentrating on one feature,First, for each strong classifier we set a default value for its strength, and then keep adding weak classifier until reaching a preset intensity, this process is also known as AdaBoost training. From the composition we can see that a Cascade classifier is a kind of decision tree structure. AdaBoost decision tree structure illustrated with Figure (8).

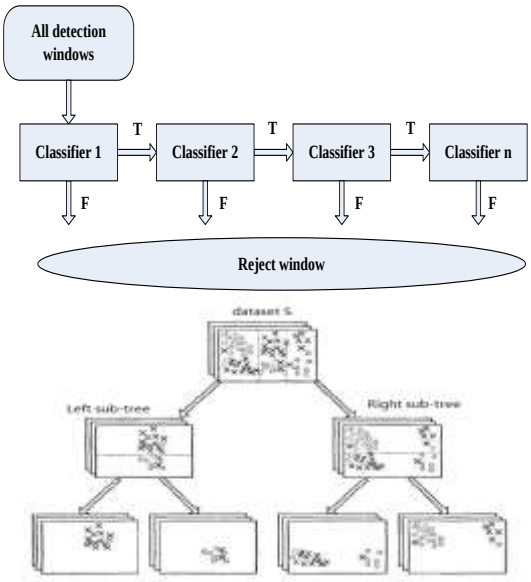


Figure 7. The Structure of the Cascade Classifier Figure 8. AdaBoost Decision Tree Structure

We use the algorithm described [19] to construct screening-cascade classifier, each feature is mapped into a 36-dimensional vector to characterize the block,each cascade of weak classifiers generated by the linear SVM training, Furthermore, Due to all the features in the detection window block training will definitely take a very long time, we have introduced a sampling algorithm Scholkopf and Smola [20] proposed, in their paper mentioned that only after several attempts, you can find the maximum value in all the random variable, they also noted that in order to find the best assessment with 95% probability in existing assessment, a random sub-sample should be $\log(0.95)/\log(0.05) \approx 59$, it can be as much as possible to ensure that we can find the best one from all the random variables, in practice, we choose 250 samples.

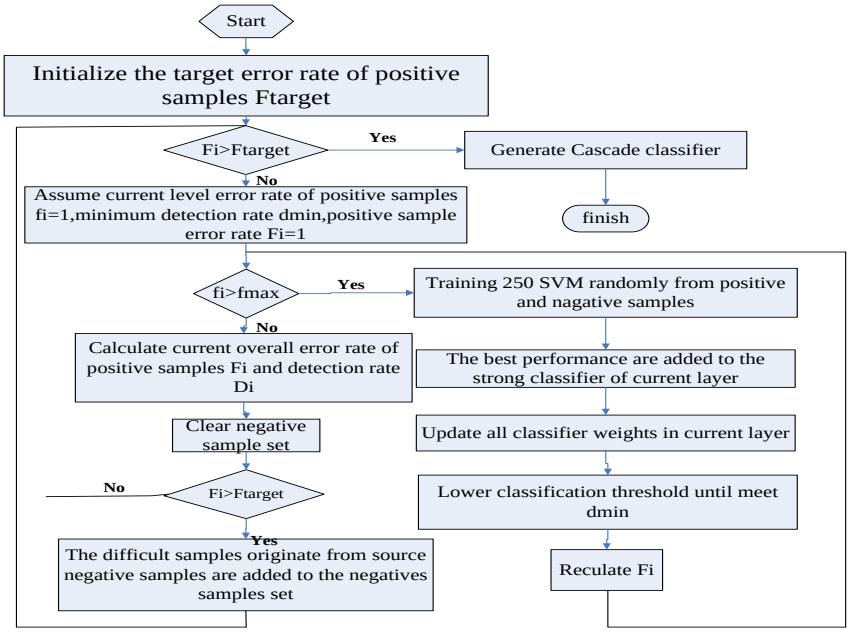


Figure 9. The Complete Training Algorithm Flowchart based on Cascade Classifier

4.1. Analysis of the Detection Result

Figure (10), summarizes the training of Cascade structure, from the figure it can be judged as the Cascade levels increasing the difficulty of training is also increasing, the number of weak classifiers from the initial 4-8 to 30, it can be concluded that weak classifiers in initial layers have a significant effect on classification capacity of the entire system. the increase in classification speed is also precisely because of classification capacity of these layers.

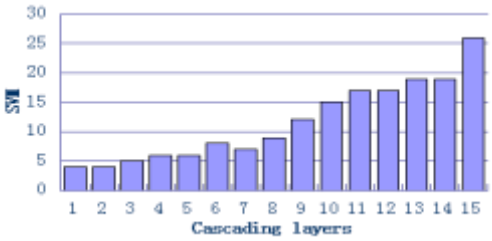


Figure 10. The Training Result of Cascade Classifier

5. Non-Maximum Convergence

When intensive scanning is carried out on windows, scanning windows are usually overlapping and have different sizes so that one object of an image usually has multiple overlapping test results, and the results must be merged together. Based on a certain rule, the non-maximal convergence algorithm singles out a final solution from an array of test results, with the remaining results being discarded.

The algorithm is based on two premises as follows:

(1) If the detector is strong enough, it should be able to make a strong positive reaction for a detect target, *i.e.*, it is able to circle the target object in the image correctly. As is often the case, the result frame is allowed to be slightly off the target center or is different from the size of the optimal result frame.

(2) A trusted detector should not react in the same frequency and confidence parameters of the target area to detect non-target area of the image. There are enough obvious differences between them.

5.1. The Principle of the Non-Maximum Convergence Algorithm

Under the two premises, the principle of the algorithm can be simply interpreted: when a bandwidth is given, all solutions within the bandwidth (*i.e.*, test results) will have a maximum, and each solution converges to the maximum at their own speed according to their own confidence parameters and frequency of other surrounding solutions. In the whole image, when many objects exist, there are theoretically many maximums which are the final test results, or the final solution of the maximum convergence algorithm.set $y_i, i = 1 \cdots n$ as the three dimensional space， S_i as the value of the scale space， H_i as a diagonal matrix， $diag[H]$ as three diagonal elements，then

$$diag[H] = [\exp(si)\sigma x, (\exp(si)\sigma y)^2, (\sigma s)^2] \tag{7}$$

Where $\sigma_x, \sigma_y, \sigma_s$ are user-defined smoothing parameters， it used to limit the smooth-width, smooth-height and smooth-scale. all the detection results represent for a kernel density estimation, we can find some local maxima, those maxima is the fusion result, in this case , we use a Gaussian kernel smoothing, so the weighted kernel density estimation[21] of point y defined by:

$$f(y) = \frac{1}{n(2\pi)^{3/2}} \sum_{i=1}^{n} |H_i|^{-1/2} t(\omega_i) \exp(-\frac{D^2[y, y_i, H_i]}{2}) \tag{8}$$

with:

$$D^2[y, y_i, H_i] \equiv (y - y_i)^T H_i^{-1} (y - y_i) \tag{9}$$

Where $D^2[y, y_i, H_i]$ is Mahalanabois distance between y and y_i， $t(\omega_i)$ is a weight for each detection result. Then the gradient for (8) is formulated as:

$$\begin{aligned} \nabla f(y) &= \frac{1}{n(2\pi)^{3/2}} \sum_{i=1}^{n} |H_i|^{-1/2} H_i^{-1} (y_i - y) \exp(-\frac{D^2[y, y_i, H_i]}{2}) \\ &= \frac{1}{n(2\pi)^{3/2}} \left[\sum_{i=1}^{n} |H_i|^{-1/2} H_i^{-1} y_i t(\omega_i) \exp(-\frac{D^2[y, y_i, H_i]}{2}) \right. \\ &\quad \left. - \left\{ \sum_{i=1}^{n} |H_i|^{-1/2} H_i^{-1} t(\omega_i) \exp(-\frac{D^2[y, y_i, H_i]}{2}) \right\} y \right] \end{aligned} \tag{10}$$

Let ϖ_i be a normalized weight of each detection result:

$$\varpi = \frac{|H_i|^{-1/2} H_i^{-1} t(\omega_i) \exp(-\frac{D^2[y, y_i, H_i]}{2})}{\sum_{i=1}^{n} |H_i|^{-1/2} H_i^{-1} t(\omega_i) \exp(-\frac{D^2[y, y_i, H_i]}{2})} \tag{11}$$

And meets $\sum_{i=1}^{n} \varpi = 1$. (10) divided (11) we can obtain :

$$\frac{\nabla f(y)}{f(y)} = \sum_{i=1}^{n} \varpi_i(y) H_i^{-1} y_i - \left(\sum_{i=1}^{n} \varpi_i(y) H_i^{-1} \right) y \tag{12}$$

let

$$\boldsymbol{H}_h^{-1}(\boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{\varpi}_i(\boldsymbol{y}) \boldsymbol{H}_i^{-1} \tag{13}$$

Where(13)is the weighted average of bandwidth matrix H_i at point y, according to the formula (12) and (13), variable-bandwidth Meanshift vector is defined as following:

$$\begin{aligned} m(y) &= H_h \frac{\nabla f(y)}{f(y)} \\ &\equiv H_h \left[\sum_{i=1}^{n} \varpi_i(y) H_i^{-1} y_i \right] - y \end{aligned} \tag{14}$$

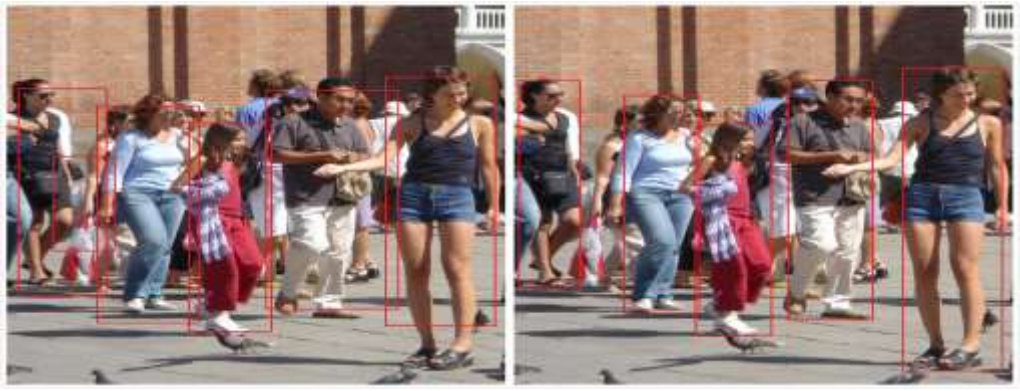
The detection result at congervence point should meet it gradient of weighted kernel density estimation is 0, then detection result convergence formela should be:

$$y_m = H_h(y_m) \left[\sum_{i=1}^{n} \varpi_i(y) H_i^{-1} y_i \right] \tag{15}$$

Applying (15) to calculate at point y_i，until y_m no longer change，then we can get a convergence position，merge multiple overlapping detection result by using formula (15) and (13)[48], and that y_m is the fusion estimation.

5.2. The Implementation of Non-Maximum Convergence Algorithm

From the Figure (11), we can clearly see that before fusion there be a heavy overlapping box, However, After using Non-maximum convergence algorithm the detection result is clearer



a. before fusion b. after fusion

Figure 11. Comparison of Fusion Result

6. Experiments

In the following, we give a detailed analysis of our approach compared to the state-of-the-art in multi-object tracking.

6.1. Evaluation Metric

We use the widely accept CLEAR performance metrics, Multi-object Tracking Accuracy (MOTA) and Precision(MOTP). Summarize performance, MOTP describe the average distance between tracks and true positives, lower values mean better precision, MOTA describe the fraction of errors, higher accuracy scores are better.

Also, we adopt the evaluation metric defined in [22],including:

Recall (↑): correctly matched detection / total detection in ground truth.

Precision (↑): correctly matched detection / total detection in the tracking result

MT (↑): the ratio of mostly tracked trajectories, which are successfully tracked for more than 80%

ML(↓): the ratio of mostly lost trajectories, which are successfully tracked for less than 80%

Table1. Performance of the Proposed Method, OLDAMs and PRIMPT. MOTA: Accuracy, Higher is Better. MOTP: Precision(cm), Lower is Better

method	MOTA	MOTP	Recall	Precision	MT	ML
Proposed method	**0.94**	**13**	**82.5%**	**88.4%**	**78.8%**	**3.7%**
Proposed method without template matching	0.85	21	80.3%	85.6%	76.6%	5.4%
Proposed method without N-mC algorithm	0.83	24	78.9%	84.8%	77.5%	5.6%
OLDAMs[23]	0.91	15	80.4%	86.1%	76.8%	4.3%
PRIMPT[24]	0.88	17	79.3%	86.6%	77.2%	5.1%

The best MOTA and MOTP scores are highlighted in bold

From Table 1, we see that by using the proposed approach the overall performance is better than the up-to-date approach, However, only by using both template matching and Non-maximun convergence algorithm our method shows excellent performance, when not in using template matching and Non-maximum convergence algorithm the tracking performance is not as good as OLDAMs and PRIMPT, Primarily because template can solve illumination change and reformed posture effectively, while Non-maximum convergence algorithm improve tracking accuracy significantly and reduce the number of missing targets.

7. Conclusions

In this paper, we proposed a multi-object tracking algorithm based on HOG template matching and Non-maximum convergence, the HOG features based human detection method could detect human of variant poses in complex background,then use the appropriate classifier to train them, Finally complete the object detection and recognition tasks. Besides, All methods are implemented in C++, OpenCV-2.4.9 and Matlab 2014a, Experiments were performed on an Intel 3.2GHz CPU and 4GB RAM.

Acknowledgement

This work was financially supported by the Natural Science Foundation of China (No.61272077), The National Natural Science Foundation of China(No.61272077), Key Laboratory Open Foundation of Jiangxi province(No.TX201204005), and Innovation Fund Designated for Graduate Students of Jiangxi Province(YC2015037).

References

[1] C. Steger, M. Ulrich and C. Wiedemann, “Machine Vision Algorithms and Applications”, Weinheim: Wiley-VCH, **(2008)**, pp. 1.
[2] W. C. Holton, "By Any Other Name". Vision Systems Design 15 (10). ISSN 1089-3709, Retrieved (2013) **(2010)** October.
[3] R. Lienhart and J. Maydt, “An extended set of Haar-like features for rapid object detection", ICIP02,

(2002), pp. 900–903.
[4] Viola and Jones, “Rapid object detection using a boosted cascade of simple features”, Computer Vision and Pattern Recognition, **(2001)**.
[5] Z. Y. A. Zhu, “P-pack SVM: Parallel Primal gradient descent Kernel SVM”, ICDM, **(2009)**.
[6] C. W. Hsu and C. J. Lin, "A Comparison of Methods for Multi-class Support Vector Machines". IEEE Transactions on Neural Networks, **(2002)**.
[7] T. Menzies and Y. Hu, “Data Mining for Very Busy People”, IEEE Computer, **(2003)**, pp. 18-25.
[8] Z. H. Deng, Z. Wang and J. Jiang, “A New Algorithm for Fast Mining Frequent itemsets Using N-Lists”, SCIENCE CHINA Information Sciences, vol. 55, no. 9, **(2012)**, pp. 2008 – 2030.
[9] P. Zhou, W. Ye and Q. Wang, “An Improved Canny Algorithm for Edge Detection”, Journal of Computational Information Systems, vol. 7, no. 5, **(2011)**, pp. 1516-1523.
[10] B. T. Moeslund, “Image and Video Processing”, **(2008)**.
[11] CTU-IIG Czech Technical University in Prague, Industrial Informatics Group, **(2015)**.
[12] D. Tarditi, S. Puri and J. Oglesby, "Accelerator: using data parallelism to program GPUs for general-purpose uses", ACM, **(2006)**.
[13] D. Merrill, “Allocation-oriented Algorithm Design with Application to GPU Computing”, Ph.D. dissertation, Department of Computer Science, University of Virginia, **(2011)** December.
[14] [17] B. Wu and R. Nevatia, “Detection and tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgetet based Part Detectors”, International Journal of computer Vision, vol. 75, no. 2, **(2007)**, pp. 47-266.
[15] [18] D. M. Gavrila and V. Philomin, “Real-Time Object Detection for” Smart” Vehicles”, IEEE International Conference on Computer Vision, no. 1, **(1999)**, pp. 87-93.
[16] [19] D. Lowe, “Distinctive image features from scale invariant key points”, International Journal of Computer Vision, vol. 60, no. 2, **(2004)**, pp. 91-110.
[17] [20] B. Scholkpf and A. Smola, “Learning with Kernals”, Canbridge,MA,USA:The MIT Press, **(2002)**, pp. 145-178.
[18] [21] D. Comaniciu, “An algorithm for data-driven bandwidth selection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 2, **(2003)**, pp. 281-288.
[19] [22] Y. Li, C. Huang and R. Nevatia, “Learning to associate: Hybrid-boosted multi-target tracker for crowded scene”, In CVPR, no. 6, **(2009)**.
[20] [23] C. H. Kuo, C. Huang and R. Nevatia, “Multi-target tracking by online learned discriminative appearance models”, In CVPR, 1, 2,4, 5, 6, 7, **(2010)**.
[21] [24] C. H. Kuo and R. Nevatia, “How does person identity recognition help multi-person tracking?”, In CVPR, **(2011)**.

